Virtual Immunohistochemistry Staining with Dual-Aligned Multi-Task Feature Guidance

Supplementary Material

1. Detailed Experiment Setup

1.1. Dataset Details

The BCI dataset consists of 3,896 H&E-HER2 training pairs and 977 testing pairs. The MIST dataset includes four biomarkers, each biomarker has 1,000 testing pairs, while the number of training pairs varies by biomarker: 4,642 for HER2, 4,153 for ER, 4,139 for PR, and 4,361 for Ki67. All images in both BCI and MIST are at a resolution of 1024×1024 pixels.

1.2. Implementation

Our framework is implemented with PyTorch on a NVIDIA GeForce RTX 4090 GPU. We employ a 6-block ResNet [4] generator as G_{vis} and a 5-layer PatchGAN discriminator [6] for adversarial training. We adopt a 9-block ResNet generator as M_{hr} and M_{ir} to perform the reconstruction auxiliary tasks. The pathology classification models, M_{hc} and M_{ic}, share the same backbone as M_{hr} and M_{ir}, with a classification head added on top of the backbone. The classification head consists of a sequence of layers: two downsampling convolutional layers, a global average pooling layer, and a fully connected layer. Our framework is trained for 70 epochs. A fixed learning rate of 2×10^{-4} is used for the first 60 epochs, followed by a linear decay to zero over the last 10 epochs with batch size of 1. The hyperparameters are set as follows: $\lambda_{se} = 0.1$, $\lambda_{adv} = 2$, $\lambda_{gp} = \lambda_g = 15$, $\lambda_{std} = 0.01$. In SEL, the downsampling ratio is $r = 16$, and the margin m and temperature τ used in L_{mcl} are set to 1.2 and 0.2, respectively.

2. Evaluation on Downstream Task

To further evaluate our framework, we conducted a positive–negative classification task on the BCI dataset. We split the BCI test set into a train-subset (732 pairs) and a test-subset (245 pairs). IHC scores of 0 and 1+ are treated as negative, while 2+ and 3+ are treated as positive. However, because the IHC scores in the BCI dataset are annotated at the WSI level, they can be noisy when applied directly to image patches. To obtain a more reliable assessment of classification performance, we removed images with incorrect labels, resulting in a refined test-subset containing 188 paired images. Then, we train a ResNet50 and apply it to evaluate the virtual staining image related to the cleaned test sub-set. The results are summarized in Table 1. Our method achieves the highest ACC among all virtual staining approaches, together with the best F1 score and AUC. This demonstrates that our virtual IHC images preserve the discriminative cues necessary for downstream classification. Notably, its performance is also close to that of real IHC images, indicating the high fidelity of the generated stains.

Table 1. Classification performance comparison on BCI dataset.

Method	ACC↑	F1↑	AUC↑
CUT[8]	0.6436	0.7564	0.6067
StegoGAN[13]	0.7500	0.8459	0.7163
BiBBDM[14]	0.6543	0.7111	0.7655
PyPix2pix[7]	0.7979	0.8716	0.8137
TDKStain[10]	0.7553	0.8321	0.7582
PSPStain[1]	0.7500	0.8508	0.8494
SIM-GAN[3]	0.7394	0.8393	0.6783
Ours	**0.8085**	**0.8808**	**0.8647**
Real IHC	0.8191	0.8844	0.9174

3. Training Details

3.1. Auxiliary Model Training

Image Reconstruction Model Training. We employ self-supervised image reconstruction as an auxiliary task to train the auxiliary model to focus on fine-grained visual details. Specifically, both the H&E and IHC reconstruction models, M_{hr} and M_{ir}, are optimized using the L2 loss to encourage an accurate reconstruction of the input images:

$$\mathcal{L}_{\text{rec}} = \|M_r(I_{in}) - I_{in}\|_2^2, \tag{1}$$

where $M_r \in \{M_{hr}, M_{ir}\}$ denotes the reconstruction model and I_{in} is the corresponding input image. We separately train M_{hr} and M_{ir} on the BCI and MIST datasets. During the feature-guidance stage, we use models trained on its corresponding dataset to extract reconstruction features for guidance.

Pathology Classification Model Training. H&E (M_{hc}) and IHC (M_{ic}) classification models are introduced to provide global semantic information for feature guidance. To equip the models to extract semantic features, we train them on BCI dataset, where H&E image and its paired IHC image are annotated with IHC level (0, 1+, 2+, 3+). Although the IHC level indicates the category of the WSI from which the H&E-IHC pair was derived, we still treat it as the label for H&E and IHC images, since deep learning models can learn meaningful semantic features from noisy labels [11, 12]. During training, both the H&E classification

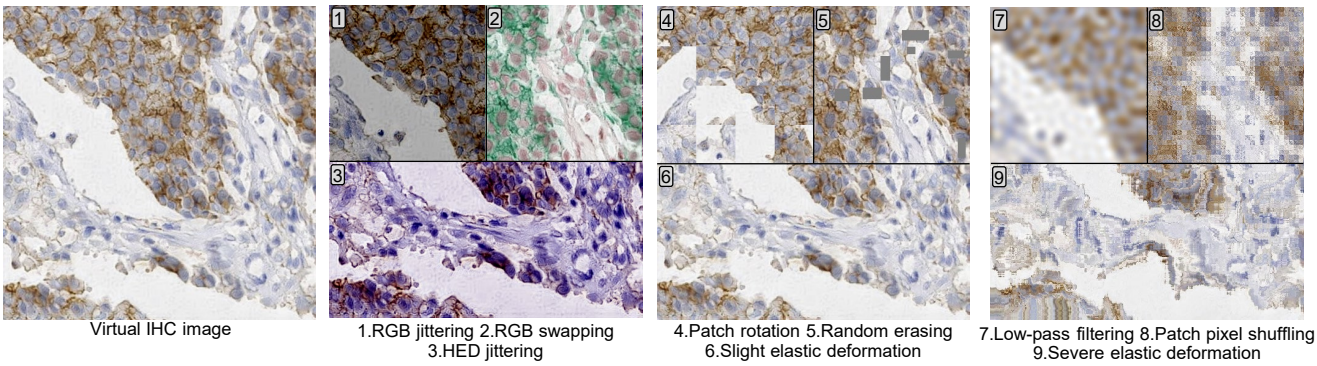

Figure 1. Visualizations of sample transformations. The structure-retained transformation is a random combination of augmentations 1–6, while the stain-retained transformation is a random combination of augmentations 7–9.

model M_{hc} and the IHC classification model M_{ic} are optimized using the standard cross-entropy loss on their staining domain. Given an input image I_{in} and its corresponding IHC level label $l \in \{0, 1+, 2+, 3+\}$, the objective is to minimize:

$$\mathcal{L}_{\text{ce}} = -\sum_{c=1}^{C} \mathbb{1}_{[l=c]} \log p_c(I_{in}), \tag{2}$$

where $C = 4$ is the number of IHC levels, and $p_c(I_{in})$ denotes the predicted probability for class c. In feature guidance, the classification models M_{hc} and M_{ic} are used to extract global semantics from both BCI and MIST image pairs, since MIST lacks IHC level annotations.

3.2. Constructing Comparative Learning Samples

In the Section 3.1 of the main paper, we implement SEL based on contrastive learning to enhance structural information in virtual IHC features. To construct contrastive samples, we design a structure-retained transformation for positive samples and a stain-retained transformation for negative ones. As shown in Figure 1, the structure-retained transformation randomly selects a subset of augmentations 1-6 to perturb staining, whereas the stain-retained transformation randomly combines augmentations 7-9 to destroy structural information.

3.3. Feature Guidance Training

In the Section 3.3 of the main paper, we describe feature guidance training. Here, we provide more details about the training process. We first extract multi-task features at two levels from auxiliary task models at different layers. At the first level of the H&E multi-task feature extraction, we obtain features from the outputs of the second ResNet blocks of M_{hr} and M_{hc}, which are then concatenated to form $F_{x_{m1}}$. At the second level, features from the outputs of their respective last ResNet blocks are concatenated to form $F_{x_{m2}}$. The IHC multi-task features are extracted in the same manner as the H&E features. Specifically, we obtain features from both the second and last ResNet blocks of M_{ir} and M_{ic}, and concatenate them to form $F_{y_{m1}}$ and $F_{y_{m2}}$, respectively. As shown in Figure 2(a), after updating M_{sf}, we obtain spatially aligned multi-task IHC features using $F_{y_{m1}}$, $F_{y_{m2}}$, and the spatial alignment matrix $\mathbf{A}$. Subsequently, for each level of the multi-task features, we construct and train a task-gap alignment model to address task-specific discrepancies.

Following the completion of step (a), step (b) focuses on leveraging the multi-task features. With M_{ta1}, M_{ta2}, and M_{sf} frozen, dual-aligned features are generated and injected into G'_{vis}. From G'_{vis}, we extract multi-scale features at three levels, which serve as semantic guidance for training G_{vis}. The auxiliary task models, M_{hr}, M_{hc}, M_{ir}, M_{ic}, are kept frozen during feature guidance training.

4. More Ablation Study Results

4.1. Supplementary Study of Feature Guidance and Dual-alignment

In the Section 4.3 of the main paper, we evaluate the effect of feature guidance and dual-alignment strategy on MIST-HER2. Here, we conduct ablation experiments on other biomarkers to further investigate their impact on virtual staining. The results are shown Table 2. Similar to MIST-HER2 results, removing feature guidance leads to a performance decline across other biomarkers. Notably, directly incorporating multi-task features or applying only spatial alignment leads to performance that is even inferior to the baseline, due to the introduction of noise during training. In contrast, incorporating task gap alignment alone improves upon the baseline, as it eliminates the task gap and mitigates spatial misalignment to some extent. These experiments further emphasize the importance of jointly addressing both spatial and task gaps to fully realize the benefits of

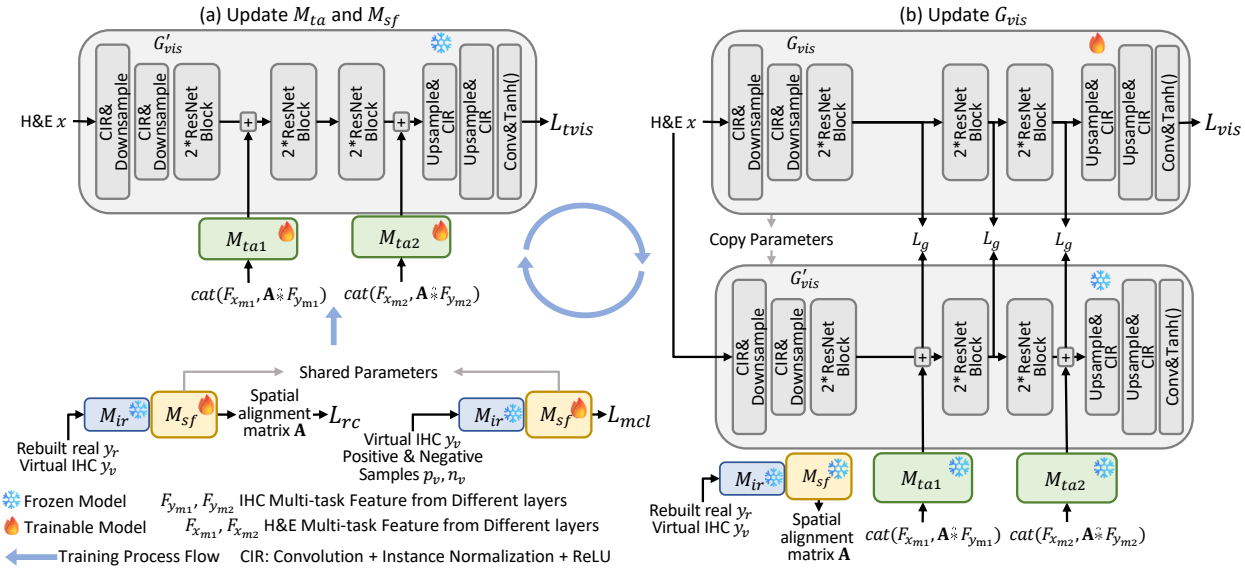

Figure 2. Feature guidance training process. The process consists of two steps: (a) Updating M_{ta} and M_{sf} to eliminate spatial and task misalignment in the multi-task features; (b) Leveraging the dual-aligned multi-task features to guide the training of G_{vis}. For clarity, the guidance loss L_g is illustrated separately from L_{vis} in step (b).

Table 2. Verification of our designed modules, where FG, SA, and TA denote feature guidance, spatial alignment, and task gap alignments, respectively. The best results are in **bold**.

FG	SA	TA	MIST-ER				MIST-PR				MIST-Ki67			
			FID↓	KID↓	LPIPS↓	SSIM	FID↓	KID↓	LPIPS↓	SSIM	FID↓	KID↓	LPIPS↓	SSIM
✘	✘	✘	39.57	10.11	0.6047	0.1980	41.22	10.67	0.5990	0.2024	34.08	9.63	0.6012	0.2268
✔	✘	✘	88.22	59.32	0.6102	0.1900	86.23	57.22	0.6081	0.2005	79.20	63.08	0.6101	0.2227
✔	✔	✘	91.91	59.33	0.6167	0.1948	72.82	35.77	0.6166	0.1927	47.18	23.20	0.6046	0.2127
✔	✘	✔	41.61	12.36	0.5995	0.2056	38.66	9.19	0.6014	0.1962	31.21	5.42	0.5993	**0.2320**
✔	✔	✔	**35.90**	**7.06**	**0.5955**	**0.2062**	**35.40**	**5.05**	**0.5975**	**0.2144**	**28.51**	**4.03**	**0.5987**	0.2222

feature guidance in virtual IHC staining. The visual comparison is shown in Figure 3, where our full framework, combining both spatial and task gap alignments, achieves the most consistent staining results, closely matching the real IHC images.

As discussed in the main text, the SEL and APM modules play a crucial role in spatial alignment. To further illustrate their impact, we provide additional ablation visualizations in Figure 4, which complement the quantitative results in Table 4 of the main paper. The visual results clearly demonstrate the importance of both SEL and APM modules. When either module is removed, the virtual staining results exhibit noticeable artifacts. These observations further confirm that both SEL and APM are essential for building reliable correspondence between virtual and real IHC images in VIS training.

4.2. Influence of Auxiliary Task Quantity

We report quantitative results on the impact of the number and type of auxiliary tasks in Table 5 of the main paper. In this subsection, we provide visualizations to further assess their effects on the visual outcomes. As shown in Figure 5, without the guidance of H&E feature, the VIS model struggles to extract meaningful information from H&E images, which in turn generates results with inaccurate staining. When we remove the rebuilt feature from the guidance, the model shows limited ability to extract detailed structures, which can be observed in the third column of Figure 5, where some nuclei are missing from the output.

4.3. Subjective Evaluation

We conducted a visual Turing test with a board-certified pathologist to evaluate the perceptual realism of the virtual IHC images generated on the MIST dataset. During the test, we randomly selected 50 real and 50 virtual IHC images from each biomarker, resulting in 400 images in total. These images were presented in a randomized order, and the pathologist was asked to classify each as either real or virtual. As reported in Table 3, 79.5% (159/200) of the virtual IHC images were misclassified as real by the expert pathol-

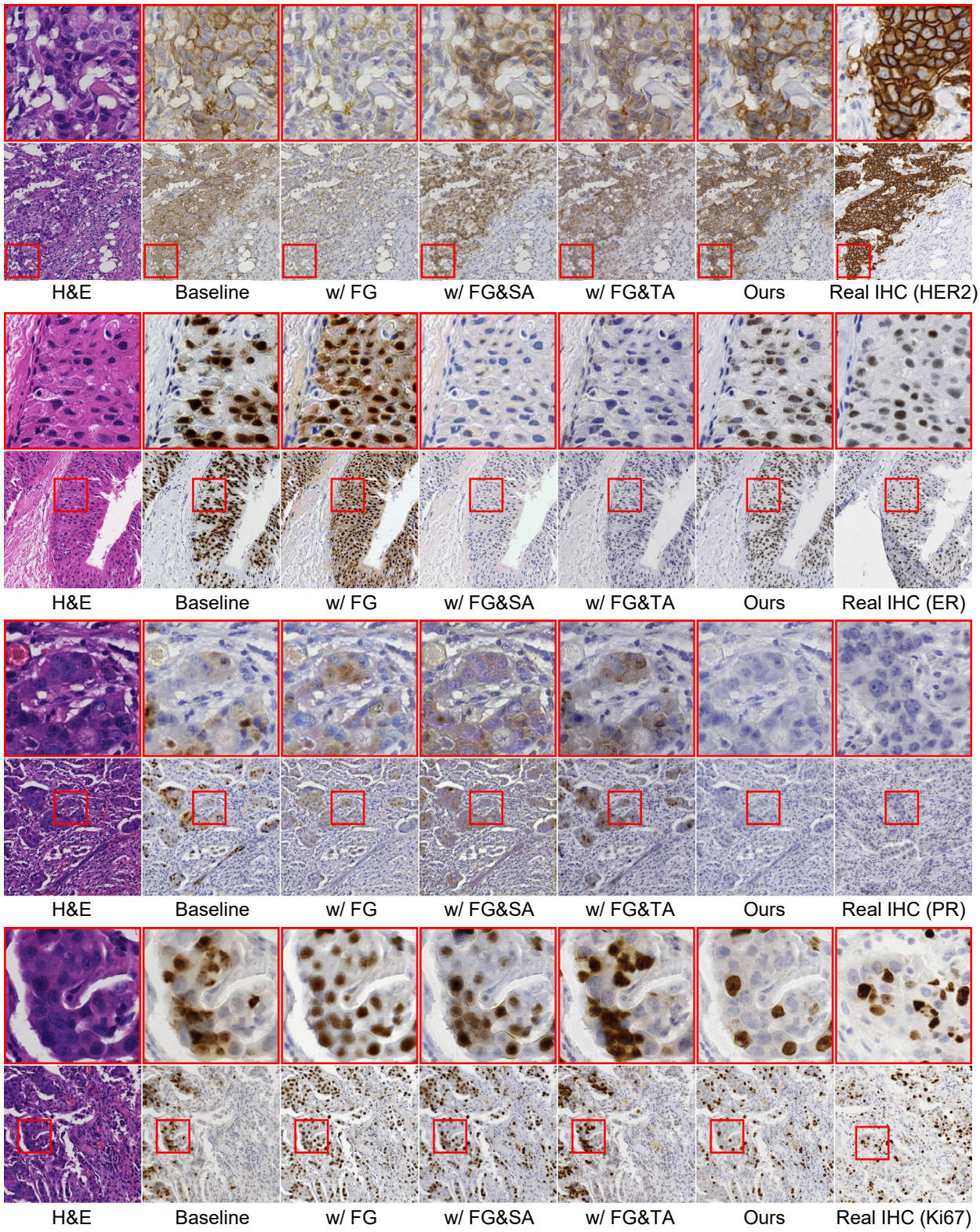

Figure 3. Visualization results of different structural variants of our framework. FG, SA, and TA denote feature guidance, spatial alignment, and task gap alignments, respectively. The baseline is built by removing FG, SA, and TA from our framework.

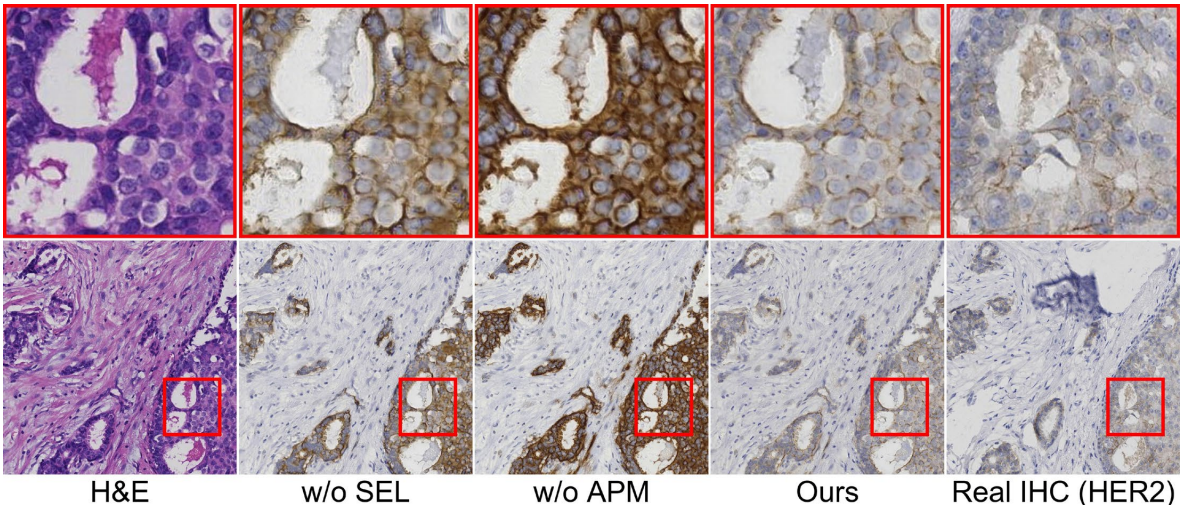

Figure 4. Visualization results of different variants of spatial alignment.

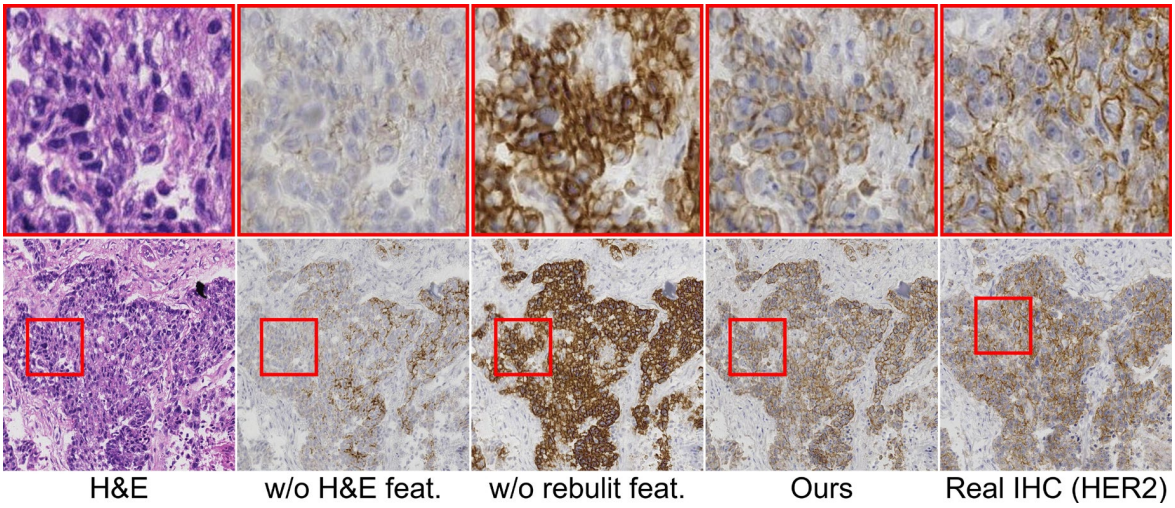

Figure 5. Visualization results of different variants of auxiliary tasks.

ogist. This high misclassification rate highlights the visual fidelity of our synthesized images and further validates the robustness and effectiveness of our method in generating realistic virtual IHC staining.

Beyond perceptual realism, we further conducted a study to directly assess whether the generated IHC images retain diagnostic information. Specifically, we randomly selected 50 ER and 50 PR sample pairs, each consisting of a generated image and its corresponding real IHC image. A board-certified pathologist was then asked to classify the biomarker expression (ER or PR) in each image as either positive or negative. As summarized in Table 4, 79% (79/100) of the generated images yielded the same classification result as their corresponding real IHC images. This high level of agreement indicates that the generated images retain clinically meaningful features and demonstrate strong potential to support accurate diagnostic decision-making.

4.4. Study on the Number of Clustering Centers K

In our work, we choose K=3 based on the histological priors, as IHC images typically contain three categories: background, negative, and positive regions. To empirically validate this choice, we conducted ablation experiments with different K values. The results are shown in Table 5. When the number of clustering centers is smaller (K=2) than the actual number of semantic categories, the model is forced to merge distinct semantics into the same cluster, which leads to ambiguous semantics of cluster centers. As a re-

sult, these centers affect subsequent matching and region similarity computation, leading to a decline in model performance. When we set K=4, it introduces more flexibility in clustering and generates clustering centers with more fine-grained semantics, which leads to a slight performance improvement. However, compared to the performance gain observed when increasing K from 2 to 3, the improvement from 3 to 4 is less substantial. One possible explanation is that the additional cluster captures a fine-grained subcategory within an existing category, for example, splitting the positive into weak and strong positive subcategories. Nevertheless, such intra-category variation is already addressed to some extent through region similarity computations in later stages, thereby limiting the incremental benefit. Overall, K=3 offers a balanced and efficient solution.

Table 3. Subjective evaluation of perceptual realism.

	HER2	ER	PR	Ki67	Total number
Real as real	48	48	47	46	189
Real as virtual	2	2	3	4	11
Virtual as virtual	13	7	7	14	41
Virtual as real	37	43	43	36	159

Table 4. Subjective evaluation of clinical correctness.

Category	Disagree Count	Agree Count	Total
Generated ER	12	38	50
Generated PR	9	41	50

Table 5. Ablation experiments with different K values.

K	FID↓	KID↓	LPIPS↓	SSIM	Training Time/epoch↓
K=2	43.23	9.38	0.5976	0.1849	0.61 h
K=3	40.34	6.56	0.5963	0.1905	0.64 h
K=4	40.23	5.99	0.5955	0.1853	0.68 h

4.5. Ablation Study of Granularity Parameter r

To investigate the effect of the alignment granularity parameter r, we performed an ablation study with $r \in 32, 16, 8$, and the results are summarized in Table 6. As shown, $r = 16$ achieves the best trade-off between performance and training cost. Larger values lead to coarser alignment and degraded image quality, whereas smaller values slightly improve performance at the expense of significantly increased training time. The setting adopted in our experiments provides a balanced choice for $20\times$ magnification images. However, we note that this value is not universal. Intuitively, the same physical region appears larger in higher-magnification images and smaller in lower-magnification ones. Therefore, we recommend adjusting r according to image resolution by using a larger r for higher magnification to accelerate training, and a smaller r for lower magnification to enhance alignment precision.

Table 6. Ablation study of granularity parameter r.

Setting	FID↓	KID↓	LPIPS↓	SSIM	Training Time/epoch↓
$r = 32$	42.85	7.95	0.6018	0.1719	0.57h
$r = 16$	40.34	6.56	0.5963	0.1905	0.64h
$r = 8$	39.95	5.95	0.5962	0.1855	0.86h

Table 7. Results of different methods on EMPaCT-p53.

Methods	FID↓	KID↓	LPIPS↓	SSIM
PSPStain	35.63	17.37	0.5839	**0.4967**
Ours	**30.08**	**14.32**	**0.5824**	0.4574

4.6. Evaluation on Different Tissue

To assess the generalizability of our method to other cancer types, we conducted an experiment on a subset of the prostate cancer dataset EMPaCT [9], comparing our method with PSPStain [1]. Specifically, we used 1,900 H&E-p53 pairs for training and 1,000 pairs for evaluation. For PSPStain, we used the same training set and followed the hyperparameter settings provided by the authors. As shown in Table 7, our method outperforms PSPStain on most evaluation metrics. These results indicate that our method exhibits generalization performance across tissue types.

4.7. Analyzing Structure and Staining Reliability.

During the early-stage training of CUT [8], we monitor grayscale SSIM for structural fidelity and cosine similarity on DAB histograms for staining fidelity. As shown in Figure 6, the structure (blue) stabilizes early, while the staining (orange) fluctuates more, supporting the claim that structure is more reliable than staining in training.

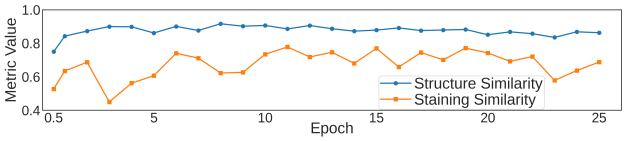

Figure 6. Convergence of Structure vs. Staining Similarity.

4.8. Analyzing SSIM on Misalignment Data

In virtual IHC staining, model performance evaluation often relies on datasets in which the input H&E and the reference IHC images are misaligned. Since SSIM operates in the pixel domain, it may be more affected by misalignment

than feature-based metrics such as FID, KID, or LPIPS. We perform a simple experiment to explore this possibility.

Specifically, we evaluate SSIM under two controlled distortions. The first, which we term shift-only distortion, involves a rigid 10-pixel shift along x axes. This operation preserves perceptual content but introduces spatial misalignment. The second, perceptual distortion, applies a combination of heavy Gaussian blurring, JPEG compression, gamma correction, and local pixelation to produce visible artifacts while keeping the image aligned with the original. The examples are shown in Figure 7. We apply these two distortions to the MIST-HER2 test set and evaluate SSIM by comparing distorted images to original images, in order to assess the SSIM's sensitivity to spatial shifts versus perceptual. The results are reported in Table 8. The shift-only distortion yields much better perceptual metric scores (FID, KID, LPIPS) than the perceptual distortion, while it receives a worse SSIM score. SSIM produces counterintuitive results, assigning worse scores to slightly shifted but perceptually intact images, and better scores to visibly distorted yet aligned ones. This suggests that SSIM may be less reliable for evaluating model performance in virtual IHC staining.

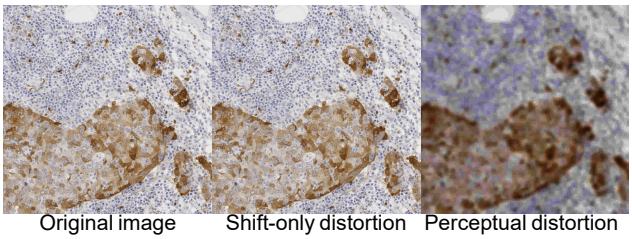

Figure 7. Visualization results of different distortions.

Table 8. Comparison of similarity metrics under spatial shift and perceptual distortion

Images	FID↓	KID↓	LPIPS↓	SSIM
Shift-only distortion	**10.91**	**0.72**	**0.3715**	0.2309
Perceptual distortion	245.89	224.11	0.7431	**0.3577**

4.9. OT-based Framework

We build an OT-based framework by replacing our spatial alignment component with OT to compare their effects on virtual staining in Section 4.3 of the main paper. Here, we describe this OT-based framework in detail.

We compute an L_2 distance-based cost matrix $\mathbf{C} \in \mathbb{R}_+^{N\times N}$ between the rebuilt real IHC feature $F_r = M_{ir}(y_r) \in \mathbb{R}^{C\times H\times W}$ and the virtual IHC feature $F_v = M_{ir}(y_v) \in \mathbb{R}^{C\times H\times W}$, where $N = H \times W$. The goal of OT is to find a transport plan $\mathbf{P} \in \mathbb{R}_+^{N\times N}$ that minimizes the global transport cost:

$$\min_{\mathbf{P}} \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbf{C}^{i,j}\mathbf{P}^{i,j} \quad \text{s.t. } \mathbf{P}\vec{\mathbf{1}} = 1, \mathbf{P}^T\vec{\mathbf{1}} = 1, \tag{3}$$

where $\mathbf{C}^{i,j}$ and $\mathbf{P}^{i,j}$ are the (i,j)-th elements of the cost matrix $\mathbf{C}$ and transport matrix $\mathbf{P}$, respectively; $\vec{\mathbf{1}}$ denotes an all-ones vector. We solve this optimization problem using the Sinkhorn-Knopp algorithm [2]. The resulting transport matrix $\mathbf{P}$ is then used to match F_{y_m} with F_v. Specifically, we reshape F_{y_m} to $\mathbb{R}^{N\times C}$, apply the transport via $F'_{y_m} = \mathbf{P}^{\mathrm{T}}F_{y_m}$, and reshape F'_{y_m} back to $\mathbb{R}^{N\times H\times W}$. The transformed feature F'_{y_m}, together with the H&E multi-task feature F_{x_m}, is used as input to task gap alignment step.

4.10. Spatial Alignment Matrix Visualization

To intuitively demonstrate how the proposed model establishes spatial correspondences between the H&E and IHC images, we visualize our spatial alignment matrices. For better visualization, a 256×256 patch is cropped to compute the spatial alignment matrix. Each alignment matrix is visualized as a heatmap, where every element represents the similarity between a spatial region in the H&E image and its corresponding region in the IHC image. To improve visual clarity, each row of the alignment matrix is normalized using min-max normalization. As shown in Figure 8, the proposed model successfully aligns each H&E image patch with its semantically corresponding IHC patch. For example, the H&E patch highlighted by the red box was originally aligned with the blue-boxed region, which is spatially adjacent but semantically inconsistent. In contrast, our alignment matrix correctly associates the red-boxed H&E patch with the semantically corresponding IHC region, as indicated by the red-marked element in the matrix.

5. More Qualitative Comparison with SOTA Methods

In Section 4.2 of the main paper, we compare our method with seven image translation approaches: CUT [8], StegoGAN [13], BiBBDM [14], PyPix2pix [7], TDKStain [10], PSPStain [1], and SIM-GAN [3]. Among these, PyPix2pix and TDKStain are based on Pix2Pix [6], StegoGAN adopts a CycleGAN-based framework [15], PSPStain and SIM-GAN are both CUT-like methods, while BiBBDM is based on diffusion model [5]. Here, we provide more qualitative results in Figures 9 and 10. Although all methods approximate the distribution of real IHC images, notable differences emerge in visual fidelity as illustrated in the red-boxed regions. In contrast, our method yields more anatomically faithful results, better preserving tissue morphology, and accurately reproducing critical biomarker expression patterns.

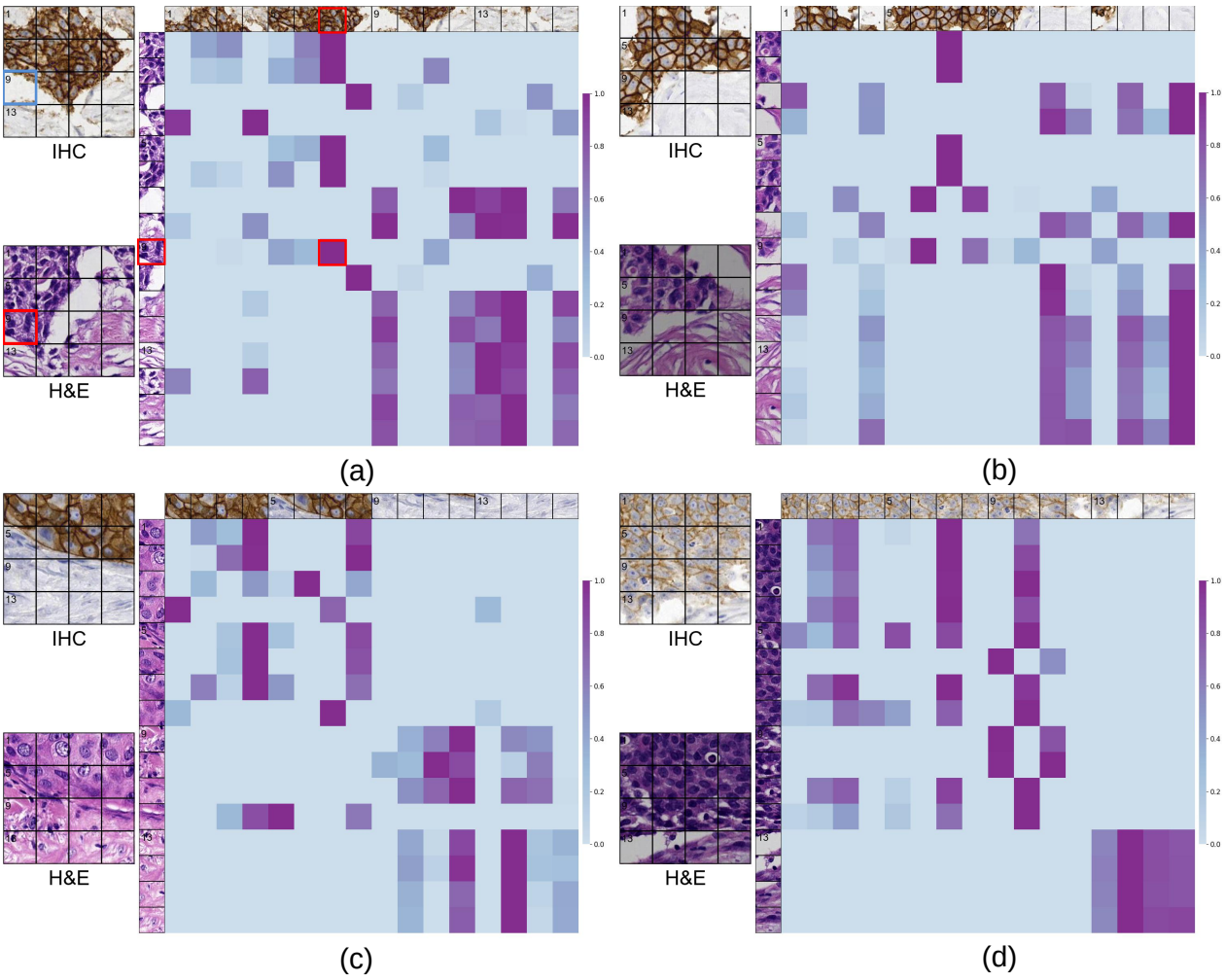

Figure 8. Visualization of spatial alignment matrix $\mathbf{A}$ on MIST-HER2 images (256×256).

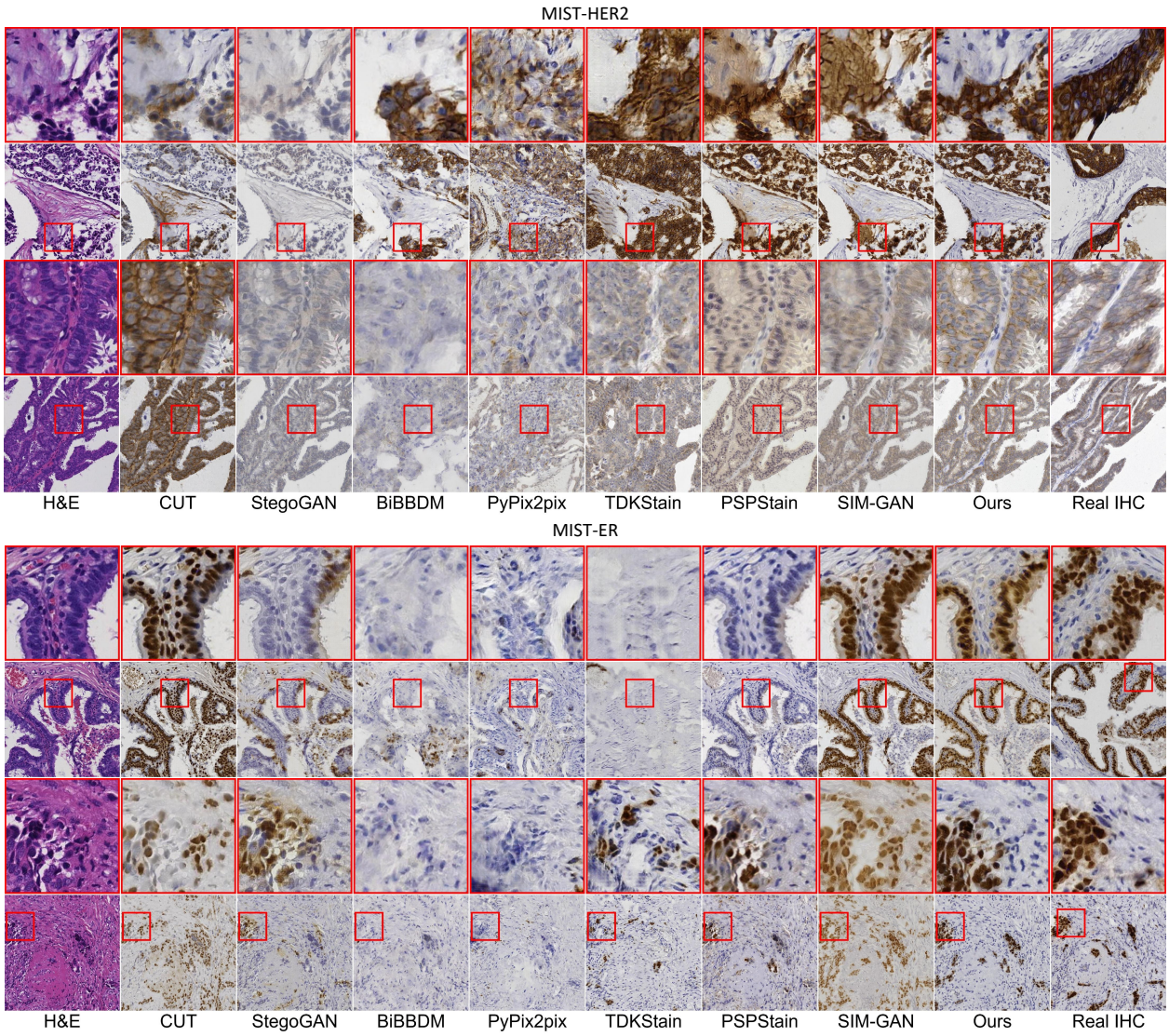

Figure 9. Qualitative results of VIS methods on MIST-HER2 and MIST-ER.

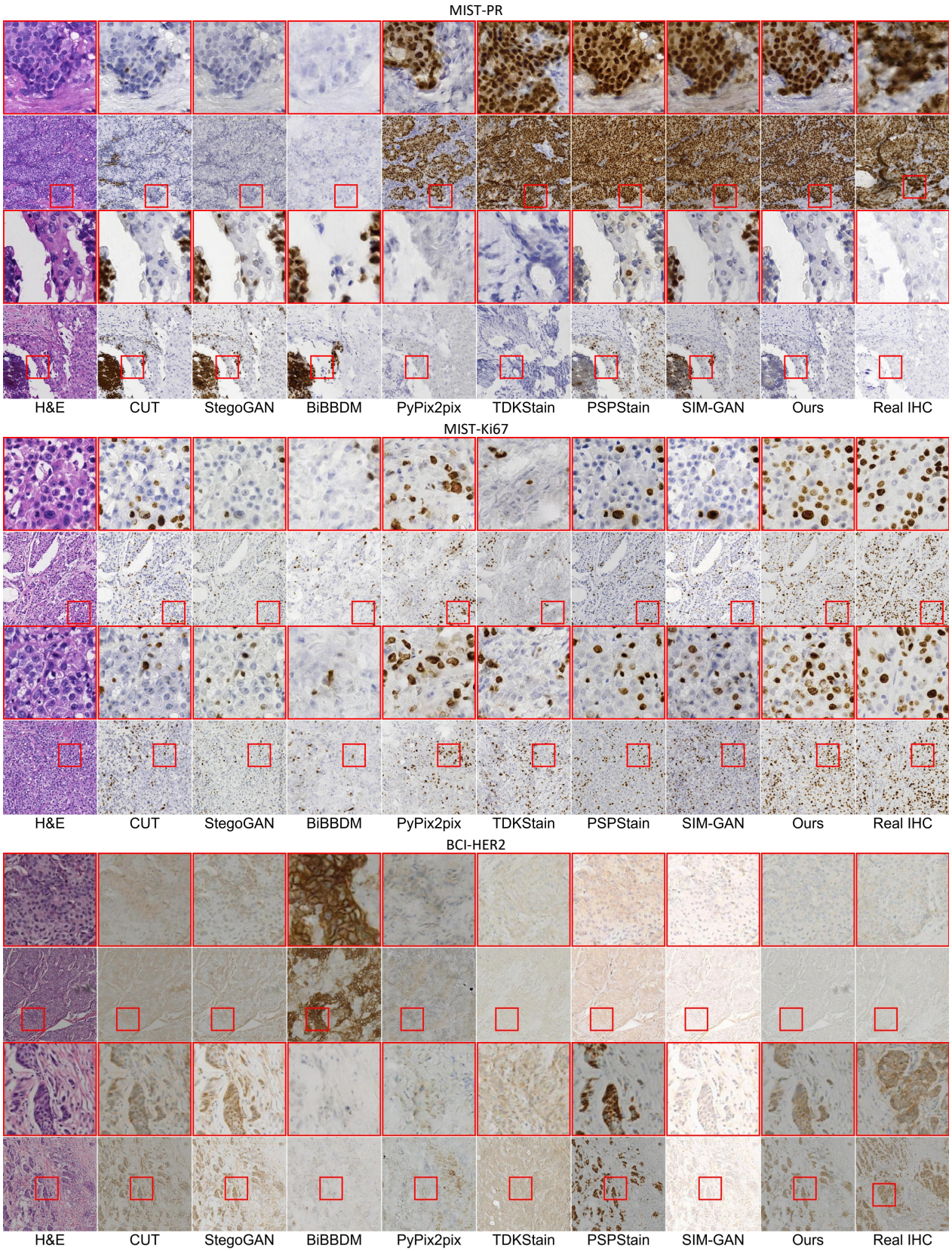

Figure 10. Qualitative results of VIS methods on MIST-PR, MIST-Ki67 and BCI.

References

[1] Fuqiang Chen, Ranran Zhang, Boyun Zheng, Yiwen Sun, Jiahui He, and Wenjian Qin. Pathological semantics-preserving learning for h&e-to-ihc virtual staining. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 384–394. Springer, 2024. 1, 6, 7

[2] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 7

[3] Xianchao Guan, Zheng Zhang, Yifeng Wang, Yueheng Li, and Yongbing Zhang. Supervised information mining from weakly paired images for breast ihc virtual staining. *IEEE Transactions on Medical Imaging*, 2025. 1, 7

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 7

[6] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1, 7

[7] Shengjie Liu, Chuang Zhu, Feng Xu, Xinyu Jia, Zhongyue Shi, and Mulan Jin. Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1815–1824, 2022. 1, 7

[8] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pages 319–345. Springer, 2020. 1, 6, 7

[9] Pushpak Pati, Sofia Karkampouna, Francesco Bonollo, Eva Compérat, Martina Radić, Martin Spahn, Adriano Martinelli, Martin Wartenberg, Marianna Kruithof-de Julio, and Marianna Rapsomaniki. Accelerating histopathology workflows with generative ai-based virtually multiplexed tumour profiling. *Nature machine intelligence*, 6(9):1077–1093, 2024. 6

[10] Qiong Peng, Weiping Lin, Yihuang Hu, Ailisi Bao, Chenyu Lian, Weiwei Wei, Meng Yue, Jingxin Liu, Lequan Yu, and Liansheng Wang. Advancing h&e-to-ihc virtual staining with task-specific domain knowledge for her2 scoring. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–13. Springer, 2024. 1, 7

[11] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017. 1

[12] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135–8153, 2022. 1

[13] Sidi Wu, Yizi Chen, Samuel Mermet, Lorenz Hurni, Konrad Schindler, Nicolas Gonthier, and Loic Landrieu. Stegogan: Leveraging steganography for non-bijective image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7922–7931, 2024. 1, 7

[14] Kaitao Xue, Bo Li, Ziyi Liu, Zhifen He, Bin Liu, Congxuan Zhang, and Yu-Kun Lai. Bibbdm: Bidirectional image translation with brownian bridge diffusion models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1, 7

[15] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 7